Detection of Web Subsites: Concepts, Algorithms, and Evaluation Issues

Eduarda Mendes Rodrigues
Microsoft Research Ltd.
7 J J Thomson Avenue
Cambridge, CB3 0FB, UK
eduarda@microsoft.com

Natasa Milic-Frayling
Microsoft Research Ltd.
7 J J Thomson Avenue
Cambridge, CB3 0FB, UK
natasamf@microsoft.com

Blaz Fortuna
Dept. of Knowledge Technologies
Institute Jožef Stefan
Jamova 39,1000 Ljubljana, Slovenia
blaz.fortuna@ijs.si

Abstract

Web sites are often organized into several regions, each dedicated to a specific topic or serving a particular function. From a user's perspective, these regions typically form coherent sets of pages characterized by a distinct navigation structure and page layout—we refer to them as subsites. In this paper we propose to characterize Web site structure as a collection of subsites and devise a method for detecting subsites and entry points for subsite navigation. In our approach we use a new model for representing Web site structure called Link Structure Graph (LSG). The LSG captures a complete hyperlink structure of a Web site and models link associations reflected in the page layout. We analyze a sample of Web sites and compare the LSG based approach to commonly used statistics for Web graph analysis. We demonstrate that LSG approach reveals site properties that are beyond the reach of standard site models. Furthermore, we devise a method for evaluating the performance of subsite detection algorithms and provide evaluation guidelines.

1. Introduction

The World Wide Web includes millions of Web sites and continues to grow in size and complexity [4,17]. Hyperlinks that connect Web pages enable users to access content within and across sites. When browsing through pages of a particular site, users rely upon search, navigation menus, A-Z index or a sitemap to find relevant information. However, despite such navigation aids many users have problems orienting themselves and completing their tasks [16].

In this paper we explore a concept of a subsite as a subunit of a Web site content and structure that may offer alternative representations of the site organization. Web sites are often organized into coherent regions. Making such organization more transparent is expected to increase the efficiency of users' search and navigation strategies. However, there are several challenges. First, we need to define a notion of a subsite that is sufficiently general, easily understood, useful to the users and, at the same time, feasible to compute. Second, we need to design algorithms for identifying subsites that are applicable to a wide range of Web sites. Finally, we need to devise a method for evaluating subsite detection algorithms. Our research attempts to address these three points.

We start with a definition of a subsite that originates from user research [15] and investigate techniques for decomposing a site into a collection of subsites. Our unique contribution is a new Link Structure Graph (LSG) representation of a Web site. Similarly to [2,5,23], the LSG method uses a page layout analysis and identifies blocks of navigation and content links. It provides a model of site organization that captures the entire hyperlink structure and preserves information about associations of links such as navigation menus or lists of links referring to related content. The LSG representation enables us to design efficient algorithms for detecting subsites and identifying entry pages for subsite navigation.

Currently, there is no easy way for Web site authors to designate organizational units such as subsites. Consequently, there is no data readily available to evaluate subsite detection algorithms or to apply machine learning algorithms that require training data. Therefore, it is essential to take the approach that enables us to collect the users' assessments of subsite decompositions and define effective evaluation measures that help us improve the algorithms.

Manually defining the scope of each subsite is potentially prohibitive even for medium size sites. Thus we suggest focusing first on identifying *entry pages* for subsite navigation. Led by the experience of the Information Retrieval (IR) community [26], we propose a pooling method to collect entry page candidates and then engage human assessors to manually evaluate which of the candidate pages indicate distinct subsites. We illustrate this method on a couple of Web sites and use the results of our analysis to assess the approach and propose guidelines for a larger scale evaluation.

In the following sections we motivate the site structure analysis, introduce the Link Structure Graph (LSG) model, and propose a method for detecting subsites using the LSG (§§2-4). In (§5) we demonstrate

the use of LSG for analyzing a sample of Web sites. We follow by describing the evaluation procedure for subsite detection and illustrate its use on two distinct Web sites (§6). We conclude with the summary of our findings and suggestions for furthering research in subsite structure analysis (§7).

2. Motivation and Background

Over a decade ago, Nielsen pointed out the benefit of adopting a structured view of the site that consists of subsites [15]. He provided an informal definition and guidelines for designing subsites:

"*By subsite, I simply mean a collection of Web pages within a larger site that have been given a common style and a shared navigation mechanism. This collection [...] should probably have a single page that can be designated the home page of the subsite*".

Nielsen emphasizes two aspects: (1) shared navigation mechanism and consistent page style and (2) importance of providing an entry point to a subsite. Without having an exhaustive set of attributes to compare different page designs it is difficult to take that aspect into account. Thus, we focus on the *shared navigation mechanism* as a defining property of a subsite. We take the same approach when identifying entry pages for subsites, looking for pages that facilitate navigation of the subsite.

2.1. Web Graph

The structure of the Web and individual sites is typically represented as a directed graph whose nodes are Web pages and edges are hyperlinks that connect them. Properties of such graphs have been extensively analyzed [4] and used in various applications such as improving the quality of search engine results [10,18], classifying Web pages and sites [1,9], and devising effective compression algorithms for storing the Web graph [21,24].

However, the authors of Web pages organize links into groups, i.e., *link blocks*. Navigation menus, for example, are repeated across pages to provide a common browsing mechanism. Some link blocks provide access to a coherent set of pages and serve as a 'hub' to access related content. This diverse structure of hyperlinks is not reflected in the standard Web graph model while it could be exploited in the detection of subsites suggested by Nielson [15]. For that reason, we introduce a new method for representing link structure of Web sites, the Link Structure Graph (LSG).

Kumar *et al.* [12] and Qin *et al.* [20] worked on a topical characterization of Web sites by using the hierarchy derived from the URL tree, i.e., the directory structure of a site. While this approach provides an effective way of analyzing the content of a site, it does capture the hyperlink structure that is essential for site browsing.

3. Link Structure Graph

The LSG of a Web site consists of nodes that correspond to distinct link blocks identified through page layout analysis. In order to cover a wide range of Web designs we define link blocks broadly as *elements of the Web page layout that include multiple links and have distinct formatting characteristics and functions.*

The edges between LSG blocks are used to capture 'navigability' from one block to another. Once the LSG of a site is constructed we define subsites as specific patterns in the LSG structure, as described in the following section.

3.1. Concepts and Definitions

The standard Web graph is a directed graph $G=(V,E)$ where a vertex $p \in V(G)$ represents a page and an edge $e=\{p,q\} \in E(G)$ represents a hyperlink from page p to page q. The *in-degree* of a vertex p is the number of edges in E that connect other graph vertices to p. Such vertices represent the *in-neighborhood* of p. The *out-degree* of p is the number of edges in E that connect p to other vertices in V. Such vertices represent the *out-neighborhood* of p.

Extending the in-degree and out-degree notions to link blocks, we define the *in-neighborhood* of a link block as the collection of pages that contain that block. We also refer to them as *container pages*. Similarly, we define the *out-neighborhood* of a block as a collection of all pages pointed to by the links in the block. We also refer to them as *target pages* of the link block.

For two link blocks g and h, we introduce a direct LSG edge from nodes g to h if at least one target page of block g contains the block h (Figure 1). More precisely,

Definition 1. Let us denote the *target pages* of link block g as $P_g \subseteq V(G)$ and the *container pages* of h as $Q_h \subseteq V(G)$. Then there is a *directed edge* from g to h if and only if $|P_g \cap Q_h|>0$.

This definition does not imply that two connected link blocks co-occur on a page. Instead, it means that at least one link in the first block refers to a page that contains the second block.

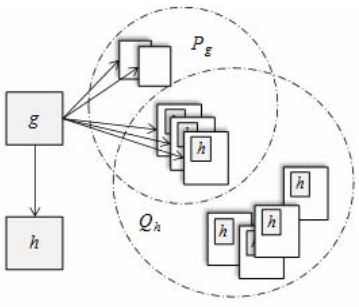

Figure 1. Edge between two LSG nodes. P_g designates target pages of the block g and Q_h represent the container pages of the block h.

The LSG model allows us to use alternative definitions of block linkage for different applications. We chose the definition that enables us to analyze navigation properties of a site. Besides blocks' connectivity, the edges of the resulting LSG capture aggregate information about links between the target pages P_g of a block g and the target pages P_h of a block h. The edge weight between the blocks is defined as the total number of hyperlinks from P_g pages to P_h pages.

3.2. Link Block Types

The LSG captures the entire set of hyperlinks on the site, grouping them into three types of structures: *structural link blocks*, *content links blocks*, and *isolated links sets*.

Definition 2. A link block comprising two or more links is a *structural link block* or *s-node* if every target page of the block contains the block itself, i.e., $P_s \subseteq Q_s$.

Definition 3. A link block comprising two or more links is a *content node* or *c*-node if $|P_c \cap Q_c|=0$. Essentially, *c*-nodes represent blocks that are not repeated across its target pages.

Definition 4. A link on the page is an *isolated link* if it does not part of any link block. For the present research we collect all the isolated links into a bag-of-links, or *i*-node.

3.3. LSG Algorithm

Our algorithm for generating LSGs involves several steps. First, we identify candidate link blocks by parsing the structure of the HTML DOM of any given page. The literature reports on successful use of the DOM structure to identify HTML layout templates [2,8] and to partition Web pages into content-coherent page units [5,28]. We follow a similar approach to extract link blocks based on HTML DOM structure. Our algorithm detects link blocks as a sequence of l or more hyperlinks that share the same common ancestor in the DOM tree structure. Here, we take a conservative approach and construct link blocks from sequences of hyperlinks only. If the list of hyperlinks is interrupted with text it is split across multiple link blocks.

Once the blocks on a page are identified, we classify each block as either *s*-node or *c*-node depending on the properties of its target and container page sets (see *Definitions 2* and *3*). The algorithm checks for repetition of each block across multiple Web pages. Given that HTML layout inconsistencies may occur, we define a threshold for the minimum number of target pages that should contain that same block in order to qualify for an *s*-node. This makes the *s*-node classification more robust in situations where the link block is not fully replicated across the pages. Once the block repetition is verified, the set of container pages is updated accordingly. Blocks that do not qualify for *s*-nodes are designated as *c*-nodes.

Isolated links that are extracted from individual pages are grouped as *bags-of-links* and classified as *i*-nodes without further processing. Finally, we create edges between blocks following *Definition 1*.

4. Segmentation of Web Sites into Subsites

In the past, researchers studied the connectivity and navigability of Web sites by analyzing connected components [19] and other connectivity properties [29] of the Web graph. However, no attempt so far has been made to incorporate navigation aspects of a Web site that are explicit in the design of Web pages and thus relevant for the user's experience. Our approach, using the LSG representation, achieves that. The LSG model captures the presence of navigation menus through *s*-nodes. Moreover, the LSG edges indicate the navigability from one menu to another. Assuming that transitions between menus indicate presence of distinct subsites, the LSG provides a good basis for segmenting a Web site into subsites.

4.1. Detecting Subsites and Entry Pages

We apply Tarjan's linear-time algorithm [25] to identify Strongly Connected Components (SCCs) of the LSG graph, i.e., maximal subsets S of LSG *s*-nodes such that any two nodes in S are reachable through directed LSG paths from one node to another. For each SCC, we define a *subsite* as the union of all the pages that contain the blocks of the SCC. This leads to regions of a Web site that are accessible through a sequence of navigation menus.

Following Nielsen's suggestion ([15], §2) we also attempt to identify suitable entry pages for subsite navigation. For that, we consider approaches that have been taken to identify quality pages on the Web.

PageRank [18] is a core link analysis algorithm for Web search and mining, which models users' navigation as a random surfing model. The rank of each page depends on the number of in-links that the page receives and the rank of the pages that contribute to the in-links. Typically, computation of the PageRank involves only in-links from external pages, assuming that they are less biased and thus a more reliable predictor of the page quality and importance. In our case we wish to combine the evidence for the entry page quality from both the LSG and the standard Web graph. Thus, for a given page we calculate three statistics:

(1) PageRank based on the Web graph—the probability that a user will navigate to a given page when randomly surfing the standard link graph G of a site:

$$PR(p_i) = \frac{1-k}{|V(G)|} + k \sum_{p_j \in N^-(p_i)} \frac{PR(p_j)}{d^+(p_j)} \qquad (1)$$

From (1) it is clear that the *PageRank* $PR(p_i)$ depends on the page rank $PR(p_j)$ and the number of out-links $d^+(p_j)$ from each page p_j in the in-neighbourhood of p_i . Here k is the damping factor and $|V(G)|$ is the number of nodes in the graph.

(2) Link block accessibility—the probability that the user will see a link block on a page considering the random surfing model on the Web graph. We calculate the *Block PageRank for* g_i as the sum of page rank scores of its container pages $Q(g_i)$:

$$BPR(g_i) = \sum_{p_j \in Q(g_i)} PR(p_j) \quad (2)$$

(3) Link block accessibility through LSG—the probability that the user will see a link block on a page if randomly surfing the pages using only the blocks that are included in the LSG. The rank of a block g_i is calculated from the ranks of all the blocks that have an edge connecting to g_i:

$$BR(g_i) = \frac{1-k}{|V(LSG)|} + k \sum_{g_j \in N^-(g_i)} \frac{BR(g_j)}{D^+(g_j)} \quad (3)$$

We combine the above measures to obtain the overall *Entry Page Rank* (*EPR*) of a page p_i for a given subsite:

$$EPR(p_i) = PR_{site}(p_i)^{\alpha} \cdot PR_{subsite}(p_i)^{\beta} \cdot BR_{site}(g_i)^{\delta} \quad (4)$$

where PR_{site} and $PR_{subsite}$ are the *PageRank* scores calculated from the full Web site graph and the subsite graph, respectively. The $BR(g_i)$ is the Block Rank of the highest ranked *s*-node included in the page p_i. The parameters α, β, and δ are defined to control the contribution of PR_{site}, $PR_{subsite}$ and BR_{site} to the overall score. They are to be determined empirically as a part of the algorithm evaluation process. Preliminary experiments suggest $\alpha=3$, $\beta=2$, and $\delta=1$.

The EPR definition reflects our intuition that a good entry page should be easily accessible from pages of the whole Web site as well as from pages within a subsite. Furthermore, good entry pages are likely to belong to a navigation menu associated with the subsite and thus receive relatively high BR scores.

5. Empirical Analysis of Site Structure

In this section we illustrate the use of the LSG for analyzing the characteristics of individual Web site structure. For that we selected a sample of 20 Web sites[1] from 7 top-level topic categories of DMOZ [7]. The sample consists of 2 to 3 sites from each category, varied in size. Table 1 presents the details of the Web site crawls. For this discussion, we selected 3 interesting aspects of Web sites, specifically related to the LSG representation.

A. Directory and hyperlink organization of sites.

Two common strategies for analyzing Web site structure involves mapping the Web site graph onto a simpler hierarchical structure, for example the tree structure inherited from the crawling strategy (breadth first, depth first, etc.) or the directory structure of the Web site reflected in the URLs of the pages [12,20]. From the Table 1 we see that these two hierarchical structures are not necessarily correlated. For example, for www.sigmaxi.org the highest percentage of pages, 36.6%, falls within depth 3 in the navigation hierarchy. At the same time, 86.9% of the pages are found within the top three levels in the directory structure.

A recent study points out that links are the most prevalent navigation element for exploring Web content [27]. Direct access to URLs, by typing the URL and moving through the directory levels, is far less common. Given that the LSG model represents a complete set of navigation elements within a Web site, the analysis based on LSG provides benefits beyond the simplified navigation tree model or the directory tree model.

B. Analysis of the LSG representation.

The LSG representation enables us to study the composition of a Web site in terms of the navigation and content links used across pages. We define two LSG statistics for a link block:

- *Block reach*: The percentage of site pages that are reached by the block, i.e., in the set of target pages.
- *Block spread*: The percentage of site pages that contain the particular block.

By definition, LSG *s*-nodes are key elements to the user's navigation within the site while *c*-nodes provide access to content pages. The larger the reach, the broader the coverage of the Web site content through the blocks. The larger the spread, the broader the replication of the link blocks across pages. For sites designed using page templates, the *s*-node spread reveals how widespread the use of a particular template is across pages in the site.

Table 1 shows that the reach and the spread of *s*-nodes and *c*-nodes vary significantly across sites. It is apparent that some sites have a wide spread of *s*-nodes and are likely to have a large number of pages sharing the same layout template. For example, menus of the www.pbs.org site target just 2.4% of site pages but are present in 98.8% of all the pages. The www.berkeley.edu site, on other hand, has various *s*-nodes blocks pointing to 23.1% of pages on the site. At the same time 78.8% of pages contain these blocks, indicating that most of the pages are well equipped with navigation menus.

Considering the reach and the spread of content nodes, it is apparent that Web sites like www.nws.noaa.gov, www.worldbank.org and www.elib.cs.berkeley.edu are content sites, where content link blocks reach between 61% and 80% of all the pages on the site. For www.nws.noaa.gov these content blocks are spread across only 2% of pages. In the www.worldbank.org case, more than 30% of pages contain content links.

It is also interesting to compare the PageRank of Web pages that include navigation menus, i.e., *s*-node containers, verse those that include content-type link blocks, i.e., *c*-node containers. We expect that many pages contain both *s*-nodes and *c*-nodes. Table 1 presents, for each site, the average BPR (eq. (2), §4.1) of

[1] These sites have been used before for analysis of Web evolution [17].

Table 1. Analysis of 20 sample Web sites.

Site	Crawl Size	%Pages (depth)	%Pages (dir level)	s-nodes		c-nodes		s-nodes	c-nodes
				Reach	Spread	Reach	Spread	BPR	BPR
1. elib.cs.berkeley.edu	855	53.1% (4)	36.1% (2)	0.056	0.058	0.614	0.151	0.254	0.746
2. eserver.org	247	36.9% (5)	59.5% (3)	0.051	0.445	0.207	0.039	0.675	0.325
3. etext.lib.virginia.edu	1805	29.2% (6)	24.4% (4)	0.213	0.650	0.316	0.148	0.852	0.148
4. nnlm.gov	4917	44.6% (6)	46.1% (5)	0.056	0.895	0.127	0.016	0.682	0.318
5. www.acsh.org	40005	37.7% (6)	40.2% (5)	0.133	0.270	0.432	0.093	0.271	0.729
6. www.artifice.com	9381	28.3% (6)	85.8% (2)	0.370	0.961	0.257	0.150	0.731	0.269
7. www.berkeley.edu	10483	46.7% (5)	53.1% (5)	0.231	0.788	0.203	0.045	0.720	0.280
8. www.biostat.wisc.edu	4251	40.0% (10)	49.4% (10)	0.006	0.024	0.529	0.197	0.514	0.486
9. www.boston.com	10380	57.8% (6)	54.2% (2)	0.444	0.389	0.125	0.029	0.778	0.222
10. www.cancerbacup.org.uk	5945	28.5% (6)	31.7% (7)	0.074	0.278	0.164	0.020	0.391	0.609
11. www.eff.org	12363	73.2% (4)	42.6% (3)	0.008	0.046	0.334	0.026	0.556	0.444
12. www.hopkins-aids.edu	18464	48.8% (5)	89.9% (2)	0.005	0.022	0.154	0.017	0.374	0.626
13. www.irs.gov	25277	32.1% (6)	68.5% (3)	0.180	0.800	0.419	0.266	0.461	0.539
14. www.nws.noaa.gov	7977	52.8% (7)	71.4% (4)	0.034	0.100	0.800	0.020	0.730	0.270
15. www.osha.gov	10939	51.9% (10)	44.6% (3)	0.201	0.683	0.305	0.083	0.748	0.252
16. www.pbs.org	3284	49.6% (10)	98% (3)	0.024	0.988	0.062	0.013	0.618	0.382
17. www.sigmaxi.org	1635	36.6% (3)	86.9% (3)	0.264	0.772	0.471	0.064	0.650	0.350
18. www.usgs.gov	770	28.7% (4)	35.8% (2)	0.115	0.394	0.466	0.184	0.527	0.473
19. www.wdvl.com	5328	35.4% (6)	38.2% (4)	0.145	0.824	0.588	0.404	0.780	0.220
20. www.worldbank.org	19607	43.7% (5)	76.0% (3)	0.098	0.424	0.655	0.306	0.519	0.481

s-node blocks (sBPR) and *c*-node blocks (cBPR). We observe that for most sites the average sBPR is higher than the average cBPR. This is expected considering that navigation blocks are repeated across pages to facilitate navigation.

C. LSG and subsite decomposition.

In the following section we focus on the method for evaluating decomposition of Web sites into subsites. Here we observe characteristics that arise from subsites detected through SCCs of the LSG (see §4.1).

The distribution of in-link degrees for the Web graph is known to approximate a power law [3,11]. Dill *et al.* [6] have shown that such property can be observed at different scales, suggesting that the Web structure has a fractal nature. We analyzed the in-degree distribution for the sample of 20 sites and observe similar long-tailed distribution (see Figure 2).

Considering the decomposition of the sites into subsites, we also analyzed the distribution of in-links that a page receives from other pages of the same subsite. For each page and each subsite that the page belongs to, we compute two statistics: the number of in-links from the subsite pages only and the number of in-links from the rest of the site, i.e., external to the subsite. Figure 3 shows the two heavy-tailed distributions, one showing internal in-links and the other external in-links. While the distributions are similar in shape they are clearly separated. The frequency of in-link degrees from within subsites is consistently higher than that of outside subsite in-links. This suggests that the identified subsites are cohesive sub-regions with well linked pages.

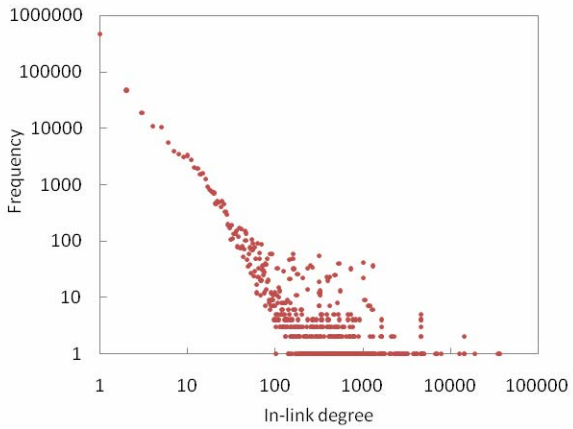

Figure 2. Logarithmic display of in-link degree distribution for all Web sites.

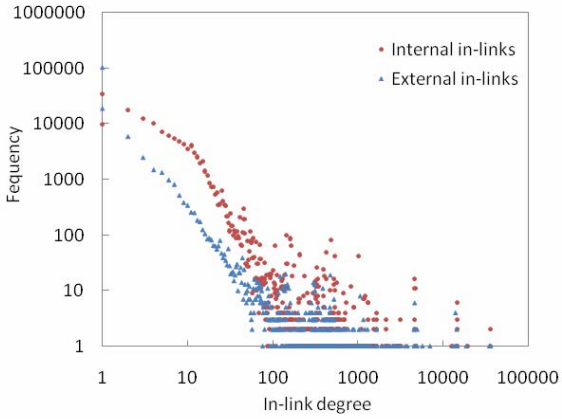

Figure 3. Logarithmic display of intra- and inter-subsite in-link degree distribution for all Web sites.

6. Evaluation Issues

Evaluation of algorithms for detecting subsites and entry pages is difficult because we do not have the correct partition of subsites for a representative set of Web sites.

Furthermore, the methods may need to be evaluated for large sites where manual inspection of all the pages is impractical. Thus we consider approaches that have been applied in other disciplines in similar situations.

In this section we propose a method for algorithm evaluation and present results of a pilot study that involves manual assessment of entry pages for subsites.

6.1. Pooling of Entry Pages

The problem of identifying all the subsites within a site is similar to the problem of obtaining an exhaustive set of documents in a database that are relevant to a given search query. For large databases it is unfeasible for users to inspect each document and render relevance judgment. The IR community approached this challenge by taking advantage of the diversity of search systems [26]. It used the pooling method to obtain candidate relevant documents from multiple systems. The top N documents from each system are contributed to the pool of documents for human judges to evaluate. In TREC [26], for example, N was 100 to make sure that the pool of unique documents exceeded the expected number of relevant documents in the database.

The problem of evaluating subsites differs from the search problem in two aspects. First, a subsite is determined by its *scope*, i.e., the set of pages that belong to the subsite, and by an *entry page* that serves as the 'home page' of the subsite and facilitates access to other pages within its scope. Thus, there are two aspects to assess as opposed to search where only the *relevance* of retrieved documents needs to be verified.

Second, evaluation of the candidate subsites and entry pages involves interaction with the site itself. This contrasts with search where each candidate document is assessed in isolation from others and without consulting the whole database. For that reason it is unlikely that an assessor of search results would identify a relevant document that is not in the pool.

In order to understand the full spectrum of issues we conducted a pilot study of subsite detection and algorithm evaluation involving two Web sites. For the sake of clarity and simplicity, we decided to focus on the detection of subsite entry pages and to leave the scope of a subsite aside for now.

Considering the unique characteristics of our problem, we modified the pooling method to collect candidate entry pages from a diverse set of algorithms and from at least one expert. We also allowed the assessors to contribute newly discovered subsites and entry pages to the final set.

6.2. Evaluation Methodology

The first step of our evaluation protocol involves browsing of the Web site by one or multiple experienced users to identify as many subsites and respective entry pages as possible. The definitions of subsite and entry page are provided to the users beforehand:

Subsite—a collection of Web pages that have been given a common style and a shared navigation mechanism. It may include a page that can be designated the 'home page' of the subsite. A subsite can also be a collection of Web pages that focus on a particular topic or function.

Entry page—a key page for accessing the content of the subsite. A subsite may have one or more entry pages.

For every identified entry page, the experts are asked to fill in a questionnaire about the proposed entry page and the subsite. These manually selected pages are added to the pool of pages nominated by alternative methods. Our final pool consisted of 5 types of entries:

A. Entry pages manually selected by experts,
B. Pages from the Web site included in DMOZ [7],
C. Index pages such as 'index.*' or 'default.*',
D. First target page of all *s*-node link blocks, and
E. Top ranked page, according to the EPR score, for each subsite detected by the LSG decomposition into strongly connected components (see §4.1).

The evaluation task consists of assessing whether the candidate pages are entry pages of subsites or not. The assessors are asked to respond *yes* or *no*, indicate the level of confidence in their assessment using the 7-point *Likert* scale, and optionally add comments. We created a simple GUI to aid evaluation—it loads each candidate page into a Web browser and collects the assessors' judgment with the confidence level of their decision.

6.3. Preliminary Evaluation Results

In this section, we report on the study findings for two Web sites: www.artifice.com and www.sigmaxi.org. The set of manually selected entry pages consisted of 5 pages for the first site and 6 pages for the second. Altogether, we had to evaluate 34 pages for the first site and 246 pages for the second site. Table 2 provides further details about the two pools of entry pages.

All the pages were judged by each of the two assessors, J1 and J2. In Table 3 we show the frequency with which they agreed and disagreed in their judgments. The observed positive agreement (*yes-yes*) was 10% for the first site and 10.4% for the second site. The observed negative agreement (*no-no*) was 72.7% and 74.9%, respectively.

Table 2. Number of entry pages detected by each method and overlap of pages by any pair of methods.

	Site: www.artifice.com					Site: www.sigmaxi.org				
	A	B	C	D	E	A	B	C	D	E
A	5	1	0	3	1	6	0	6	2	0
B		2	0	0	0		2	2	0	0
C			4	1	1			114	10	15
D				24	6				125	41
E					10					70

Table 3. Agreement scores among assessors J1 and J2.

		Assessor J1					
		Yes	No	**Total**	Yes	No	**Total**
J2	Yes	1	4	**5**	7	17	**24**
	No	5	24	**29**	43	179	**222**
	Total	**6**	**28**	**34**	**50**	**196**	**246**
		Site: www.artifice.com			Site: www.sigmaxi.org		

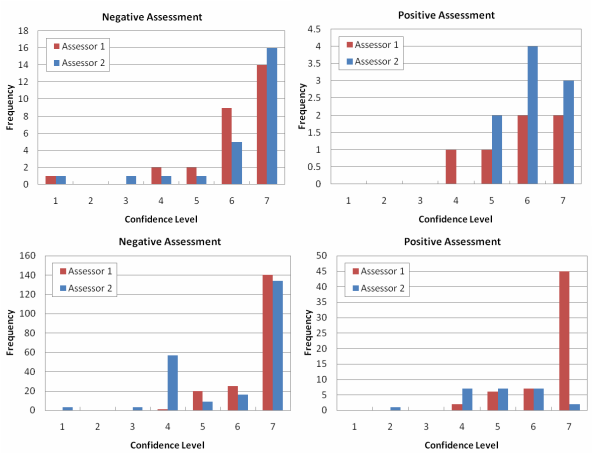

Figure 4. Histograms of the confidence (1-not confident at all, 7-very confident) on negative (left) and positive assessments (right) for www.artifice.com (top) and www.sigmaxi.org (bottom) sites.

Figure 4 relates the number of negative and positive assessments to the confidence levels declared by the assessors. Here the lower score corresponds to the lower confidence level (1='not confident at all', 7='very confident'). The statistics shows that the assessors were quite confident about their negative judgments and less so about their positive judgments.

6.3.1. Assessment of Entry Page Candidates

We illustrate the use of assessors' judgments by analyzing the quality of page types A-E (§6.2) that were included in the pool of candidate pages. Table 4 shows Set Precision and Recall statistics relative to the individual assessor's judgments.

Considering the precision statistics for manually added pages (type A), it is interesting to observe that the judges J1 and J2 agreed with the expert on 20% of suggested pages for the first site and 83% and 67%, respectively, for the second site. On this smaller set of pages, their level of agreement with the expert is higher than their mutual agreement (10%) on the full set of candidate pages. The recall statistics (R ≤17%) clearly shows that relying only on manual input would miss a significant portion of subsite entry pages.

We observe that both assessors considered 2 DMOZ pages from each site to be subsite entry pages (P=100%),

Table 4. Precision (P) and recall (R) for each input method.

	A (manual)	**B** (DMOZ)	**C** (index*)	**D** (s-node)	**E** (EPR)
	Site: www.artifice.com				
Assessor J1	P: 20% R: 17%	P: 100% R: 33%	P: 25% R: 17%	P: 4% R: 17%	P: 20% R: 33%
Assessor J2	P: 20% R: 11%	P: 100% R: 22%	P: 25% R: 11%	P: 21% R: 56%	P: 20% R: 22%
	Site: www.sigmaxi.org				
Assessor J1	P: 83% R: 8%	P: 100% R: 3%	P: 49% R: 93%	P: 10% R: 20%	P: 20% R: 22%
Assessor J2	P: 67% R: 17%	P: 100% R: 8%	P: 19% R: 92%	P: 4% R: 21%	P: 13% R: 38%

but, of course, these pages are only a fraction of all the subsite entry pages that were identified (R≤33%).

We also point out the effectiveness of the simple heuristics used in C, i.e., pattern matching on 'index' or 'default' in the URL of a page. Such pages are most likely to be home pages of subsites, which contributes to the high precision scores for this approach.

Finally, we note a significant difference in the recall of entry pages between the two sites. The second site, www.sigmaxi.org is neatly organized into a hierarchy of subsites using a common template, with most of the branching nodes indicated by 'index' page. The first site, www.artifice.com does not have a unified 'look and feel' and a common template.

In that instance the LSG method discovers subsites that would not have been retrieved by method C (33% *vs*. 17% and 22% *vs*. 11% recall) nor would have been found manually (input A). We expect this to be the main contribution of the LSG based method across a variety of Web sites.

6.4. Evaluation Issues and Guidelines

The objective of the assessment process is to arrive at a *gold standard* that could be used to refine automated subsite detection. Thus, it is in our interest to provide assessors with tools that would make them more effective and efficient. There are several ways in which we can help the assessors build a good mental model of the site organization:

- Provide quick access to the pages in the vicinity of a given page, i.e., the parent, child and sibling nodes.
- Provide visual clues such as page thumbnails of flexible size.
- Make the relationship between the URL and the links on the parent page explicit.
- Provide easy access to pages that have already been visited during evaluation. Present a navigation trail and the underlying hyperlink structure.
- Enable the assessors to customize presentation of candidate entry pages, i.e., as a sorted list, graph, etc.

In order to facilitate the comparison of assessments, it is important to record any ambiguity encountered by the assessors and the rationale for the rendered judgments. Furthermore, we should investigate methods for merging the multi-assessor judgments based on the self declared confidence levels of the assessment.

We see an opportunity to engage with Web authors and administrators by providing tools for analysis of Web site structure. Major online search engines have already adopted the XML standard for describing sitemaps in order to facilitate crawling and index update [22]. Using and expanding the existing descriptors we can reduce the need for automatic detection of subsites and focus on the structure analysis and support for navigating large sites.

7. Concluding Remarks

In this paper we consider the need for improving support for Web site navigation, particularly for large sites with complex menu structure. Existing aids such as a sitemap and A-Z pages have had a limited impact since they are not context sensitive and do not represent the complete content and navigation structure of the site. The solution is to create an adaptable system that can reveal appropriate parts of the site as needed. However, the first step is to identify organizational units that comprise the site—we refer to them as subsites.

The traditional Web link graph does not offer a sufficiently rich representation to support in depth analysis of the site structure. Thus we introduce the LSG representation that incorporates information about the menu structure and blocks of links referring to content pages on the site. LSG analysis enables us to decompose sites into subsites and identifying entry pages. We illustrate how LSG can be used to analyze properties of derived subsites and show that LSG subsites are 'coherent', having a higher distribution of in-link degree from pages within the subsite than those from the rest of the site.

Recognizing the importance of evaluating decomposition of Web sites, we devised a pooling method for gathering relevance assessments and conducted a pilot study to test the approach. The pilot study enabled us to reflect on the challenges and provide guidelines for organizing a large scale evaluation. We expect that the best strategy is to combine a community based assessment of Web site structures with the input from Web authors and administrators.

8. Acknowledgement

The authors would like to thank Gavin Smyth for implementing the subsites assessment tool and Vishwa Vinay for assisting with the evaluation procedure.

9. References

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The Connectivity Sonar: Detecting site Functionality by Structural Patterns," *Proc.* 14th ACM *Conf.* HT, Aug. 2003.

[2] Z. Bar-Yossef and S. Rajagopalan, "Template Detection via Data Mining and its Applications," *Proc.* WWW'02, pp. 580-591, May 2002.

[3] A. Barabasi, R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, pp. 509-512, 1999.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, "Graph Structure in the Web", *Proc. WWW'00*, pp. 309-320, May 2000.

[5] S. Debnath, P. Mitra and C. Lee Giles, "Automatic Extraction of Informative Blocks from Webpages," In: *Proc.* ACM SAC'2005, pp.1722-1726, 2005.

[6] S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar and A. Tomkins, "Self-similarity in the web," *ACM Transactions on Internet Technology*, vol. 2, no. 3, pp. 205-223, 2002.

[7] DMOZ – Open Directory Project: http://dmoz.org

[8] D. Gibson, K. Punera and A, Tomkins, "The Volume and Evolution of Web Page Templates," *Proc.* WWW'05, pp 830-839, May 2005.

[9] E.J. Glover, K. Tsioutsiouliklis, S. Lawrence, D.M. Pennock and G.W. Flake, "Using Web Structure for Classifying and Describing Web pages," *Proc.* WWW'02, May 2002.

[10] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proc. 9th Annual* ACM-SIAM *Symposium on Discrete Algorithms*, pp. 668-677, 1998.

[11] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, "Stochastic Models for the Web Graph," *Proc. Annual Symposium on Foundations of Computer Science*, vol. 41, pp.57-65, 2000.

[12] R. Kumar, K. Punera and A. Tomkins, "Hierarchical topic segmentation of websites," In: *Proc. 12th ACM* KDD '06, pp. 257-266, Aug. 2006.

[13] C. Lindemann and L. Littig, "Coarse-grained classification of web sites by their structural properties," *Proc. 8th ACM* WIDM '06, pp. 35-42, Nov. 2006.

[14] N. Milic-Frayling and R. Sommerer, "SmartView: Enhanced Document Viewer for Mobile Devices," *MSR Technical Report* MSR-TR-2002-114, November 2002.

[15] J. Nielsen, "Alertbox for September 1996". Available at: http://www.useit.com/alertbox/9609.html

[16] Nielsen Norman Group, "Site Map Usability: 28 design guidelines based on usability studies with people using site maps," Nielsen Norman Group Report, 2002.

[17] A. Ntoulas, J. Cho and C. Olston, "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," *Proc.* WWW'2004, May 2004.

[18] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Technical report, Stanford Digital Library Technologies Project*, 1998.

[19] V. Petricek, T. Escher, I. Cox and H. Margetts, "The Web Structure of E-Government – Developing a Methodology for Quantitative Evaluation," *Proc.* WWW'2006, May 2006.

[20] T. Qin, T.-Y. Liu, X.-D. Zhang, G. Feng and W.-Y. Ma, "Subsite Retrieval: A Novel Concept for Topic Distillation," *LNCS*, vol. 3689, pp. 388-400, 2005.

[21] S. Raghavan and H. Garcia-Molina, "Representing Web graphs", *Proc.* IEEE ICDE'03, pp. 405-416, Mar. 2003.

[22] Sitemap Protocol: http://www.sitemaps.org/

[23] R. Song, H. Liu, J.R. Wen and W.Y. Ma, "Learning Important Models for Web Page Blocks based on Layout and Content Analysis," ACM SIGKDD Explorations Newsletter, 6(2), pp. 14-23, 2004.

[24] T. Suel and J. Yuan, "Compressing the Graph Structure of the Web," *Proc. Data Compression Conference*, pp. 213-222, 2001.

[25] R.E. Tarjan, "Depth-first Search and Linear Graph Algorithms", SIAM *Journal on Computing*, vol. 1, no. 2, pp.146-160, 1972.

[26] TREC–Text Retrieval Conference: http://trec.nist.gov/

[27] H. Weinreich, H. Obendorf, E. Herder and M. Mayer, "Off the beaten tracks: exploring three aspects of web navigation," *Proc.* WWW '06, May 2006.

[28] L.Yi, B.Liu and X. Li, "Eliminating noisy information in Web pages for data mining," *Proc.* SIGKDD'03, pp. 296-305, 2003.

[29] Y. Zhang, H. Zhu and S. Greenwood, "Web site complexity metrics for measuring navigability," *Proc.* QSIC'04, Sept. 2004.